CLUSTER ANALYSIS FOR CTBT SEISMIC EVENT MONITORING

Dorthe B. Carr and Chris J. Young, Sandia National Laboratories
Richard C. Aster and Xioabing Zhang, New Mexico Institute of Mining and Technology

Sponsored by U.S. Department of Energy
Office of Nonproliferation and National Security
Office of Research and Development
Contract No. DE-AC04-94AL85000

ABSTRACT

Mines at regional distances are expected to be continuing sources of small, ambiguous events which must be correctly identified as part of the Comprehensive Nuclear-Test-Ban Treaty (CTBT) monitoring process. Many of these events are small enough that they are only seen by one or two stations, so locating them by traditional methods may be impossible or at best leads to poorly resolved parameters. To further complicate matters, these events have parametric characteristics (explosive sources, shallow depths) which make them difficult to identify as definite non-nuclear events using traditional discrimination methods. Fortunately, explosions from the same mines tend to have similar waveforms, making it possible to identify an unknown event by comparison with characteristic archived events that have been associated with specific mines.

In this study we examine the use of hierarchical cluster methods to identify groups of similar events. These methods produce dendrograms, which are tree-like structures showing the relationships between entities. Hierarchical methods are well-suited to use for event clustering because they are well documented, easy to implement, computationally cheap enough to run multiple times for a given data set, and because these methods produce results which can be readily interpreted. To aid in determining the proper threshold value for defining event families for a given dendrogram, we use cophenetic correlation (which compares a model of the similarity behavior to actual behavior), variance, and a new metric developed for this study.

Clustering methods are compared using archived regional and local distance mining blasts recorded at two sites in the western U.S. with different tectonic and instrumentation characteristics: the three-component broadband DSVS station in Pinedale, Wyoming and the short period New Mexico Tech (NMT) network in central New Mexico. Ground truth for the events comes from the mining industry and local network locations, respectively. The clustering techniques prove to be much more effective for the New Mexico data than the Wyoming data, apparently because the New Mexico mines are closer and consequently the signal to noise ratios (SNR's) for those events are higher. To verify this hypothesis we experiment with adding gaussian noise to the New Mexico data to simulate data from more distant sites. Our results suggest that clustering techniques can be very useful for identifying small anomalous events if at least one good recording is available, and that the only reliable way to improve clustering results is to process the waveforms to improve SNR. For events with good SNR that do have strong grouping, cluster analysis will reveal the inherent groupings regardless of the choice of clustering method.

Key Words: cluster analysis, discrimination, mining events

OBJECTIVE

One way to determine if an event is from a particular mine is for an analyst to compare known waveforms from that mine to the unknown waveform. This can be done by eye, and for an analyst who knows a region well, events from certain mines can be easily identified by visual inspection. However, with the large volume of data that can be expected when monitoring the Comprehensive Nuclear-Test-Ban Treaty (CTBT), we want to find a way to automate comparing unknown waveforms to archived events associated with specific mines. In this study, we investigate the use of cluster analysis techniques to facilitate this comparison. Previous seismic event studies have used cluster analysis to classify event catalogues (Isrealsson, 1990; Riviere-Barbier and Grant, 1993), but have only provided a cursory view of the richness of the discipline.

Using data from three mines in Wyoming and four mines in New Mexico, we compare different waveform processing methods and cluster analysis techniques to determine what processing parameters and cluster analysis techniques do the best job of clustering known events. To help determine which methods give the best results, we use two metrics. The first is cophenetic correlation and the second is a metric based on separation of known events from different mines and the formation of clusters of events from the same mine.

RESEARCH ACCOMPLISHED

Cluster Analysis

The term cluster analysis refers to a variety of techniques for grouping similar entities in such a way that their interrelationships are revealed. Most of the work in the area comes from taxonomy, a discipline of biology, where researchers seek to classify organisms based on sets of measurements (e.g. various measurements of types of bones). The first step in any method of cluster analysis is to quantify the similarity, or conversely, the dissimilarity between the entities to be categorized. Because the mathematical equations involved are typically given using dissimilarity, we follow that convention. A measure of dissimilarity could easily be developed for seismic waveforms (e.g. parameters could be arrival times, signal to noise ratio (SNR), etc.), but this would involve a subjective choice of the parameters which we would prefer to avoid. Instead we choose to base our measure of dissimilarity on waveform correlation, by forming the complement of the correlation coefficient. This quantity conveniently ranges from 0 to 1, but this is not necessary for cluster analysis; any measure of dissimilarity will do.

Cluster Analysis Techniques

There are many different types of cluster analysis techniques, but most fall into one of four general types (Davis, 1986): partitioning methods (e.g. factor analysis), arbitrary origin methods (e.g. K-means method), mutual similarity methods, and hierarchical clustering methods. The choice of the type to use depends on various characteristics including the user's mathematical competency, the computational resources available, and of course the manner in which the results are to be used. For this study we choose to use hierarchical cluster methods, because these methods are well-documented, easy to implement, computationally cheap enough to run multiple times for a given data set, and because these methods produce results which can be readily interpreted for our data sets (seismic events). Hierarchical clustering methods form *dendrograms*, which are tree-like structures showing the relationships between the entities. Dendrograms can be built either by division, i.e. from the top down, or agglomeration, i.e. from the roots up. Agglomerative methods are much more common and are generally computationally less expensive. In this study we use agglomerative methods.

There are many methods to agglomeratively build the dendrogram, but all of them follow the same basic process. Note that in the following the term "pair" refers to any two entities that are being joined; each of these entities can be either an original data entity or a cluster formed from data entities and/or other clusters. The generalized process is as follows:

1. find the minimum dissimilarity pair (i.e most similar) in the dissimilarity matrix
2. remove the rows and columns corresponding to each member of the pair from the matrix

3. by some means add a new row and column corresponding to the dissimilarities between the newly formed pair and all remaining entities (either original entities or clusters).
4. repeat steps #1-3 until the matrix has been reduced to 2x2 whereupon the last pairing is the only pair left and no further updates are necessary.

The trick, of course, is in step #3 where one must generate dissimilarities for the new pair with all other entities. There are several methods for doing this. In our study we tested six methods. Nearest neighbor or **single linkage** is a method based on the minimum dissimilarity, i. e. the best pair of all possible pairs. On the other end of the spectrum is the furthest neighbor or **complete linkage** method where linkage is based on the maximum dissimilarity, i.e. the worst pair. With the **group average** method, linkage is based on the average of the dissimilarities. For the **centroid** method, linkage is based on the squared Euclidean distance between the centroids of each entity. Linkage is based on the "between-group sum of squares" for two entities with the **minimum variance** method. At each stage in this method, variance within clusters is minimized with respect to variance between clusters. The last method is the **flexible** method. The dissimilarities are weighted with three constants that add to 1. The values we use are 0.625, 0.625 and -0.25.

Dendrogram Interpretation

Once the dendrogram has been formed, it can be interpreted. However, there can be many ways to group the events into different clusters and the decision is always subjective. In some cases, the groups may be obvious, but this is certainly not always the case. For complex dendrograms, the question of where to "cut the stems" of the tree-like structure is not easily answered and so we have investigated methods to aid in the decision.

Cophenetic Correlation

One method which has been proposed to help with this problem is the use of cophenetic correlation. Cophenetic correlation is the correlation between the actual dissimilarities as recorded in the original dissimilarity matrix, and the dissimilarities which can be read off of the dendrogram. In essence, this is a measure of how well the dendrogram, which is a model of the similarity behavior, models the actual behavior. Notice that the cophenetic correlation can be calculated at each step of the building of the dendrogram, "scoring" only the dissimilarities between entities which have been built into the tree to that point. Thus, one can make a simple plot of cophenetic correlation vs. pair number. Sudden decreases in the cophenetic correlation indicate that the cluster just formed has made the dendrogram less faithful to the data and thus may suggest that the decision line should lie between this cluster and the previous one (Ludwig and Reynolds, 1988). Similarly, one can look for sudden increases in variance at each cycle for a suggestion that a "bad" group has been formed. For our implementation, the variance increase at each cycle is computed as the variance of the newly formed group less the variances for each of the two groups which were combined to form the group. Both of these methods are easy to implement and we have built them into our standard dendrogram package to aid in interpretation. In general, we find the cophenetic correlation to be much more robust.

SNL Metric

In the case where the grouping is already known and the method and any waveform processing parameters are being chosen to achieve the best separation, we found that it is desirable to develop a metric to score the success of the produced dendrogram at separating the groups. We found no such metrics in the literature and so developed one of our own. Our metric rewards two properties. First, separation, which we define to be the tendency to place entities from different groups into their own clusters. Thus dendrograms which put all of the objects of a certain type in clusters which have few if any objects of other types will score well. Second, in order to promote the formation of clusters, we reward fusion, that is the tendency to have fewer clusters for each group. This constraint must be added because otherwise the unclustered original data entities will score perfectly for separation and this will always be the preferred solution.

Data

We compare waveform processing methods and group similarity calculation methods using archived regional and local distance mining blasts recorded at two sites in the western U.S. with different tectonic and instrumentation characteristics. Both sites are located within regional distances of regular mining activity, so data is readily available.

In Wyoming, data was collected with the Deployable Seismic Verification System (DSVS) installed at the Pinedale Seismic Research Facility (PSRF) near Boulder, Wyoming. DSVS records three component high frequency data (0.5-50 Hz) on a Teledyne Geotech S3 seismometer and a 24-bit digitizer. The sample rate is 200 samples per second. The data can be considered accurate up to 40 Hz and the average background noise is close to the USGS-Peterson low noise model (Carr, 1993). Three mining operations in Wyoming provided origin times and total tonnage of ripple fired events detonated at their sites in 1991 and 1992. Archived DSVS data was searched for events and 175 usable signals were found. Both timing problems and poor signal-to-noise made it difficult to find signals at DSVS. The origin times supplied often did not produce a signal at the predicted P-time at DSVS, so many events were assumed to be from a particular mine if they started close to the predicted P-time. Most events were small, because the ripple-fire technique is used to reduce ground motion and minimize damage.

Mining data in New Mexico was collected with the New Mexico Tech (NMT) network, a collection of 19 stations throughout New Mexico. All the seismometers are 1 Hz, critically damped geophones with an upper corner of about 15 Hz. The sampling rate is 100 Hz. For this study we are using station CAR, a single vertical component instrument. Data was collected over a 4 month period in 1997 when mining activity was high in western New Mexico and eastern Arizona. The events were located using the NMT network. Even though the mines are located outside the network, we feel the locations are good enough to associate specific events to each mine. And the fact that the NMT data comes from a network provides us with a means to compare our clusters, which are derived using only a single station of the network, with the locations determined from the entire network using traditional methods.

Table 1 lists characteristics of the data from the mines in Wyoming and New Mexico. Mines W1 and W2 are at similar distances from DSVS and both have poor SNR. Mines NM1 and NM3 are at similar distances from station CAR in New Mexico. Examples of signals from all of the mines are in Figure 1.

Table 1: Characteristics of Wyoming and New Mexico mines

mine	distance	azimuth	Lg-P (range)	Lg -P (ave)	SNR
W1	359 km	75 degrees	38-50 sec	43.6 sec	poor
W2	322 km	90 degrees	36-47 sec	43.0 sec	poor
W3	144 km	217 degrees	17-22 sec	20.6 sec	good
NM1	179 km	224 degrees	17-24 sec	22.6 sec	good
NM2	211 km	226 degrees	25-27 sec	26.3 sec	good
NM3	144 km	336 degrees	20-25 sec	22.9 sec	good
NM4	257 km	249 degrees	28-37 sec	31.3 sec	good

Waveform Processing

Before cluster analysis is done we process the waveforms. Six parameters, (1) time window, (2) sample rate, (3) phase type, (4) tapering, (5) Hilbert enveloping and (6) filtering are varied to see the effects on the resulting dendrograms.

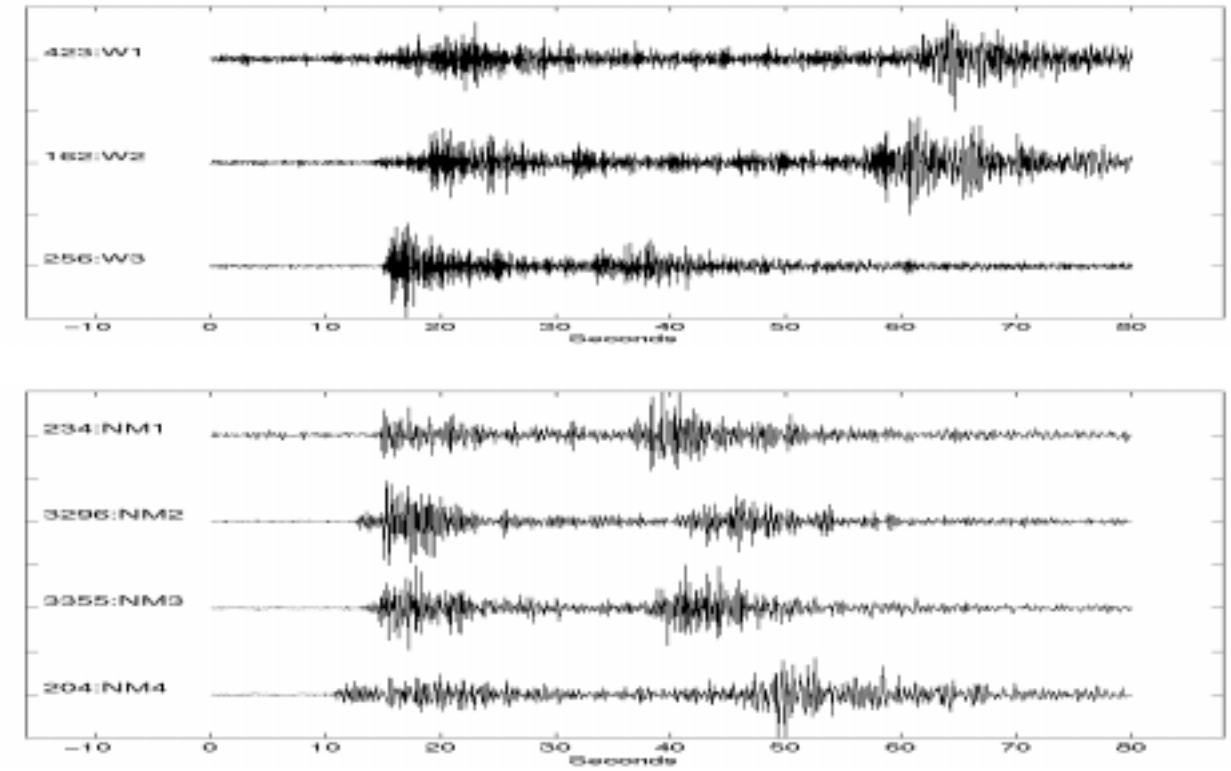

Figure 1. Sample waveforms from the Wyoming mines (top) and New Mexico mines (bottom)

Rivier-Barbier and Grant (1993) found that the Lg - P time is the characteristic of the waveform that best clustered their data, and we also found this to be true. A time window starting right before the picked P arrival and long enough to see the secondary arrival worked better than shorter windows or windows centered on the secondary arrival. The sample rate did not have much affect on the dendrograms, and so we choose to downsample the data in order to minimize the computer memory needed to calculate the dissimilarities.

Tapering the traces was done to remove any edge effects that could result from ending in the middle of a cycle. The taper cannot be too large however, or important features of the waveform will no longer contribute to the correlation. A slight taper of 5% works well. Using a Hilbert envelope removes negative values in the waveform and acts as a low pass filter. Without the envelope, the sorting focusses on small details in the waveforms that cause events from the same mine to be split into different clusters. The enveloping also causes the resemblance values to increase and separates the clusters better, so we always use a Hilbert envelope when doing cluster analysis. Filtering affects the results only if the waveforms from different mines have different frequency content. In most cases, filtering the data does not have much effect on the dendrogram.

We started by using the flexible method to do cluster analysis. Once we determined the best processing parameters using this method, we used the same parameters and tested the other five cluster analysis techniques. All methods cluster the events, although there are slight differences in the pairing of events. We found that if the waveform processing parameters are picked correctly and are robust, it really doesn't matter which cluster analysis method is used. Since we like the look of the dendrogram created using the flexible method the best, that is the method of choice.

Cluster Analysis Results

Dendrograms of the Wyoming and New Mexico data are in Figures 2 and 3. For both data sets, dendrograms are calculated using a time window starting just before the picked P arrival and long enough to see any secondary arrivals, a

5% taper, a Hilbert envelope and the flexible method. In Wyoming the events separate into three clusters, one of W3 events and two with W1 and W2 events mixed together. The decision line along the left side of the clusters is where we picked the solution, using the metrics which are illustrated beneath the dendrogram. Mines W1 and W2 are approximately the same distance from DSVS, and it appears that the Lg -P time alone cannot effectively separate the two mines into distinct clusters. We experimented with just the W1 and W2 events to see if it is possible to separate them into two clusters. Filtering the data helps, but we cannot find a characteristic feature in the W1 and W2 waveforms that clusters the two mines into two distinct groups effectively and robustly.

However, in New Mexico the events separate into four clusters, each corresponding to a specific mine. There are three events that are misclustered, but they have abnormally low P amplitudes compared to the other signals from the same mines, so we believe that is why they are misclustered. The two mines in New Mexico at similar distances from the station CAR, mines NM1 and NM3, clearly separate into two groups, unlike the Wyoming data. Mines W1 and W2 are both over 300 km away from the recording station, and the SNR is not very good. Mines NM1 and NM3 are less than 180 km away from the recording station and have good SNR. We hypothesize that since the SNR is better for the mines in New Mexico, there is correlation on other features of the waveforms besides the Lg - P time that is separating the events into two clusters. In Wyoming the SNR is not good enough to correlate on any other features in the waveforms.

In order to investigate the effects of changing SNR on the groupings provided by cluster analysis, we use the New Mexico data, which has excellent SNR in its raw form. We simulate noisy data by adding noise to the data to decrease the SNR. One common means to add noise to an event recording is to capture a sample of noise ahead of the first arrival, scale it as desired, and then add it to the event waveform. Unfortunately, for our data set this cannot always be done because the data is segmented and so each waveform does not always include a sufficient pre-event noise sample. Instead, we choose a single pre-event noise sample spanning 27 seconds from one of the event recordings, randomize the phase information to prevent correlation, and dilate/contract it as needed to generate the noise samples which were added to each of the waveforms. To do this, for each signal waveform to which we added noise, we took the fft of our master noise sample, resampled the amplitude spectra at the appropriate frequency spacing for the signal waveform's time length (recall that the length of the time interval spanned by a waveform determines the frequency discretization for the fft in the spectral domain), generated new random phase information at the new frequency discretization using a white distribution, transferred the new spectral series back to the time domain using the inverse fft to generate the new noise waveform, scaled the new noise waveform for the specified SNR, and then added it to the signal waveform. For the scaling, we use SNR to specify the relationship between the maximum value of the signal waveform and the maximum value of the generated noise waveform which will be added to the signal waveform. Thus, an SNR of 2 implies that the generated noise waveform is scaled to half of the signal waveform before it is added to the signal waveform.

We add noise to the New Mexico data in gradual steps, starting with a value of SNR = 10 and decreasing to a value of SNR = 0.5. What this means is that when SNR = 10, we added noise to the traces which has a maximum value of 0.10 of the maximum signal. Using a value of SNR = 0.5 means adding noise which has a value of 2 times the maximum signal. The dendrograms produced when SNR = 10 and SNR = 5 show no noticeable differences from the dendrogram with no noise (Figure 3). There are four clusters, each corresponding to one mine with the same misclustered events as with the raw data. Changes in the dendrogram start occurring when we get to noise levels with SNR = 3, SNR = 2 and SNR = 1. In Figure 4 we see the dendrogram resulting when SNR = 2. Sample traces without noise and with the added noise are in Figure 5. There are four clusters in the dendrogram. Mines NM2 and NM4 still make distinct clusters, but the other two clusters consist of events from both mines NM1 and NM3. These two mines are at similar distances from the station CAR (Table 1). We have added enough noise to the signals, that the characteristics in the waveforms that separated the two mines when the SNR was good can no longer be recognized by the cluster analysis. When we decrease the value of SNR to 0.5, we add so much noise that not even the Lg - P times can cluster the events, and we end up with the events from the four mines all mixed together.

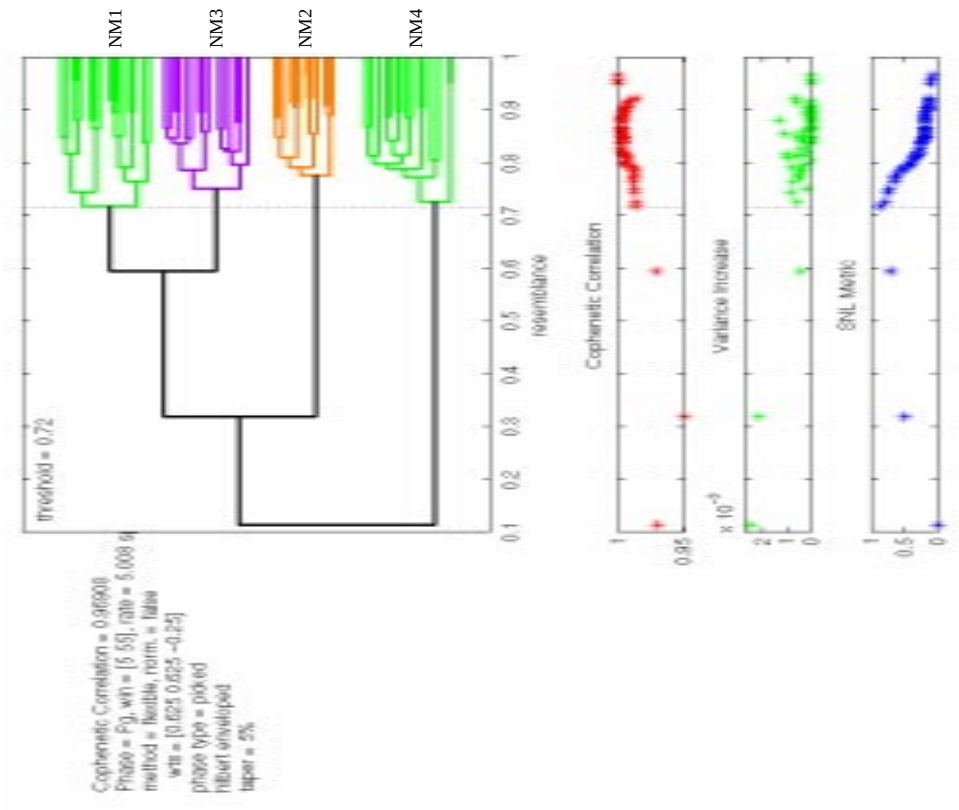

Figure 3. Dendrogram using the events from mines in New Mexico. There are four clusters corresponding to each of the four mines.

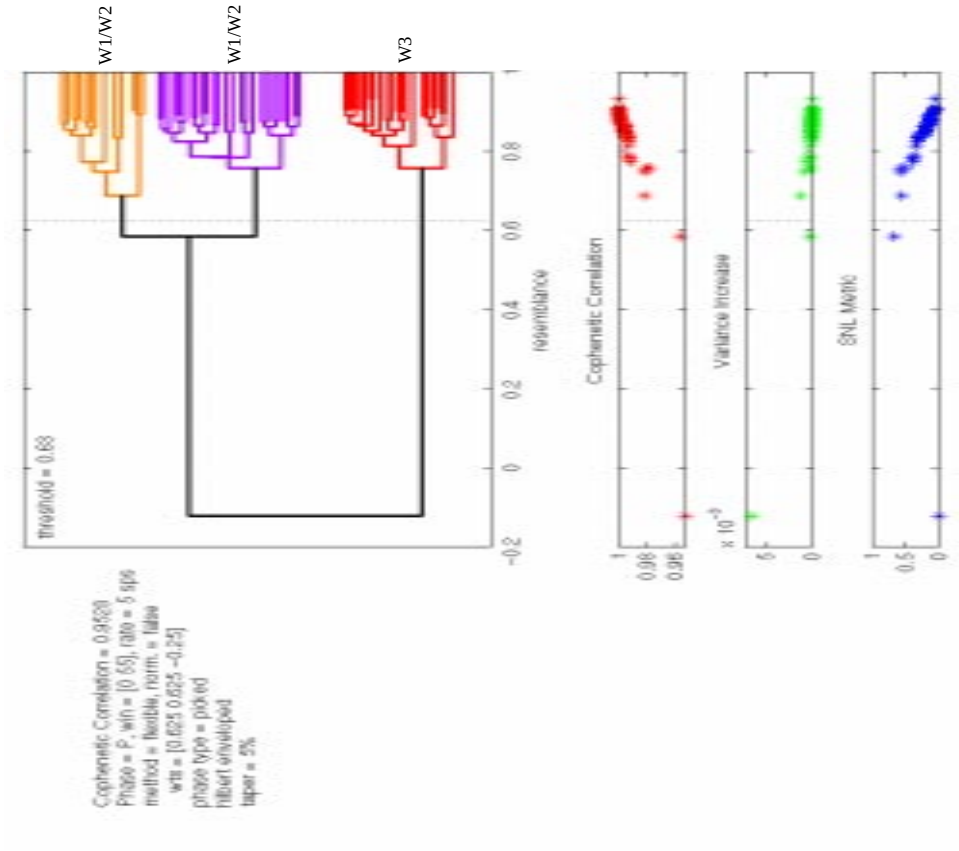

Figure 2. Dendrogram using the events from mines in Wyoming. There are three clusters, but the top two clusters both have W1 and W2 events. Only W3 events are in the bottom cluster.

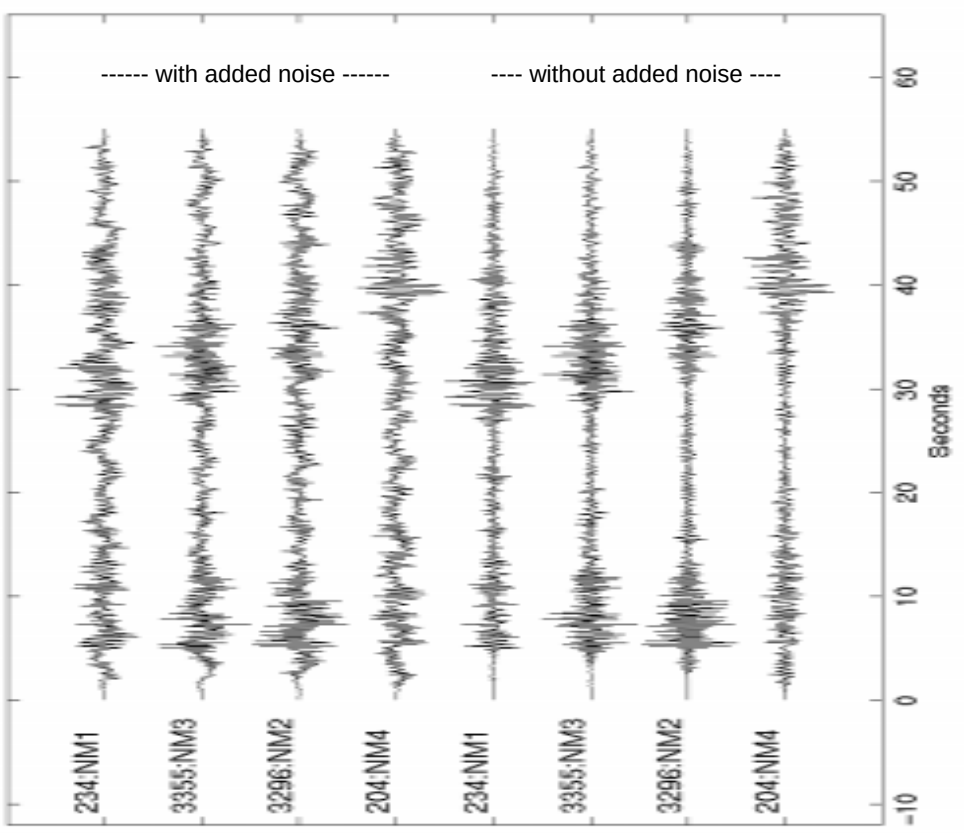

Figure 5. Sample traces from the New Mexico data. The four traces on the bottom are the raw traces, i.e. without added noise. The four traces on top are the same events where noise with a maximum value of 1/2 of the maximum signal was added to the traces.

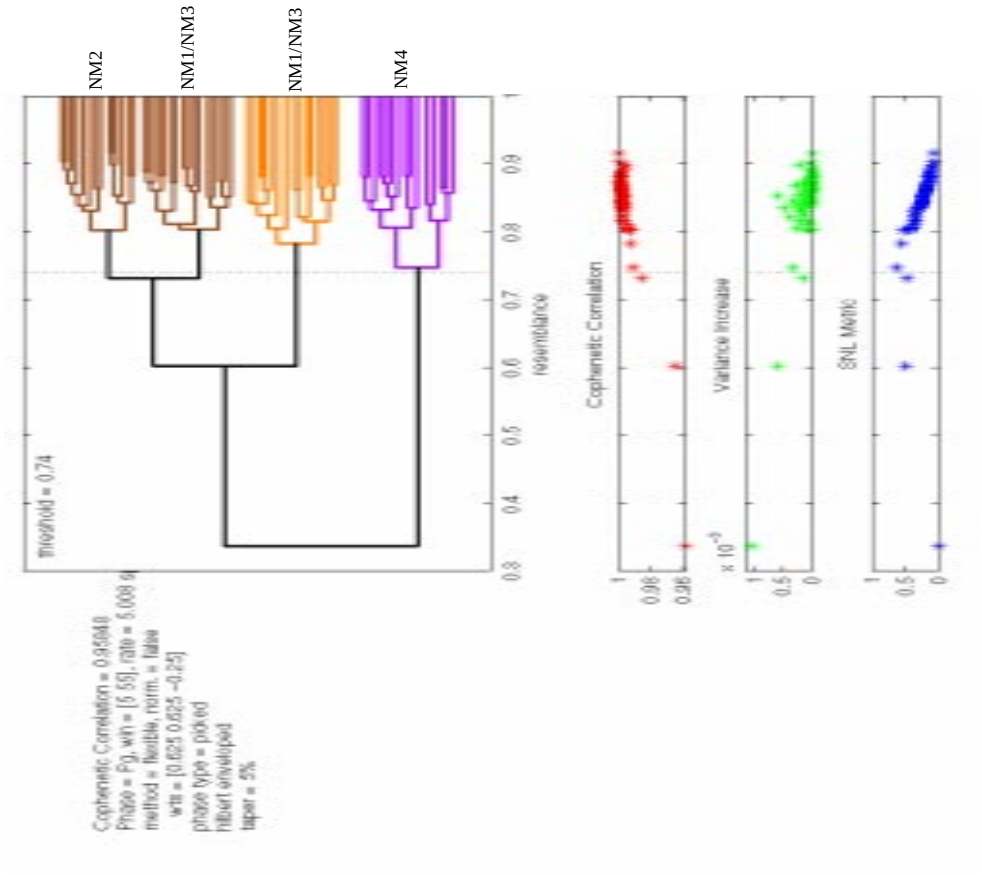

Figure 4. Dendrogram using the events from mines in New Mexico after noise was added to the raw waveforms. There are still four clusters, but the middle two both have events from mines NM1 and NM3 mixed together.

CONCLUSIONS ANDRECOMMENDATIONS

As we have shown, it is easy to calculate dendrograms for a given data set, and if appropriate waveform processing is done, events from specific mines will cluster. The waveform processing is important, and some specific parameters affect the resulting dendrogram more than others. One characteristic of the waveform that is important for creating distinct clusters is the Lg - P time, so the time window used to calculate the dendrogram must include both the P and secondary arrivals. Using a Hilbert envelope can also improve the clustering, because it smooths out some of the details in the waveforms that can cause events from the same mine to be broken into more than one cluster.

In Wyoming, no matter how much pre-processing of the waveforms we do, we are not able to clearly separate events from mines W1 and W2. We concluded at first that if the Lg - P times for two mines are virtually identical, then clustering analysis would not work. However, calculating a dendrogram with the New Mexico data proved that events from mines at similar distances can cluster in distinct groups. The difference between the Wyoming and New Mexico data is the SNR of the events. The New Mexico data had very good SNR, but the Wyoming data, for a variety of reasons, had poor SNR. By adding noise to the New Mexico data, we are able to show that clustering deteriorates as the noise level increases. When we add noise that has a maximum value of 1/2 of the maximum signal, the events from mines NM1 and NM3 which had clearly clustered with the raw data, are starting to mix together. The dendrogram from New Mexico calculated with the added noise (Figure 4) looks similar to the dendrogram calculated using the Wyoming data (Figure 2).

If the waveforms have poor SNR, then the other features that can be used to cluster events such as the shape of the different arrivals or the frequency content, cannot be discerned. In fact, the production of a dendrogram is very much a “garbage in, garbage out” process; none of the cluster analysis methods can extract information where there is none available. Typically, if an analyst cannot discern similarities between the entities, then a dendrogram will not help. However, if the preparatory work is properly done, and the data do have strong grouping, the choice of cluster analysis method will make little difference. We believe cluster analysis can be useful in comparing unknown waveforms with archived data from known mines. It should be fairly simple to implement an automated process to do the comparison.

REFERENCES

Carr, D. B. (1993). **Evaluation of the Deployable Seismic Verification System at the Pinedale Seismic Research Facility**, SAND93-1696.

Davis, J. C. (1986). **Statistics and data analysis in geology**, J. Wiley & Sons, New York.

Isrealson, H. (1990). **Correlation of Waveforms from Closely Spaced Regional Events**, Bull. Seism. Soc. Am., vol 80, pp. 2177-2193.

Ludwig, J. A. and J. F. Reynolds (1988). **Statistical ecology**, J. Wiley & Sons, New York.

Riviere-Barbier F. and L. T. Grant (1993). **Identification and Location of Closely Spaced Mining Events**, Bull. Seism. Soc. Am., vol. 83, pp. 1527-1546.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.